View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Statistics Norway's Open Research Repository

2007/9 | Rapporter | Reports

Astrid Mathiassen and Geir Øvensen

A practical approach for model-based poverty prediction

Statistisk sentralbyrå • Statistics Norway
Oslo–Kongsvinger

Rapporter

I denne serien publiseres statistiske analyser, metode- og modellbeskrivelser fra de enkelte forsknings- og statistikkområder. Også resultater av ulike enkeltundersøkelser publiseres her, oftest med utfyllende kommentarer og analyser.

Reports

This series contains statistical analyses and method and model descriptions from the various research and statistics areas. Results of various single surveys are also published here, usually with supplementary comments and analyses.

© Statistics Norway, February 2007
When using material from this publication, please give Statistics Norway as your source.

ISBN 978-82-537-7143-4 Printed version
ISBN 978-82-537-7144-1 Electronic version
ISSN 0806-2056

Subject
05.90

Design: Enzo Finger Design
Print: Statistics Norway

Symbols in tables	**Symbol**
Category not applicable	.
Data not available	..
Data not yet available	...
Not for publication	:
Nil	-
Less than 0.5 of unit employed	0
Less than 0.05 of unit employed	0.0
Provisional or preliminary figure	*
Break in the homogeneity of a vertical series	—
Break in the homogeneity of a horizontal series	|

Abstract

Astrid Mathiassen and Geir Øvensen

A practical approach for model-based poverty prediction

Reports 2007/9 • Statistics Norway 2007

The objective of this report is to provide practical guidance for producing poverty estimates based on "light" household surveys. Mathiassen (2005) outlines the theoretical model. A household budget survey is used to estimate a statistical consumption model where a small set of variables are linked to consumption and poverty. These indicators are then collected through light surveys in years where no household budget survey is made available. By combining the light survey indicators and the parameters from the consumption model, poverty rates and their standard errors can be predicted. The report takes the reader through each step of the procedure, from preparing and utilizing the survey datasets, selecting good indicators and predicting the poverty rates, to evaluating the predictions. The SPSS syntax generated by the INE workshops is available at: www.ssb.no/en/int.

Acknowledgement: This project and report are financed by The Norwegian Agency for Development Cooperation (NORAD). The authors thank Bjørn Wold and Stein Opdahl, who initiated the project and provided valuable comments along the way. We are also grateful to participants at several workshops. The first two workshops were undertaken at Instituto Nacional De Estaticia (INE) in Maputo (December 2005 and February 2006) for predicting poverty in Mozambique. The third workshop was undertaken in Oslo (October 2006) with participants from National Statistical Office (NSO) and Ministry of Economic Planning and Development (MEPD) Malawi, where we applied the method for predicting poverty in Malawi. The participants in all three workshops made a valuable contribution to the form and content of this document.

Contents

1. Introduction

The increased demand for regular and frequent monitoring of poverty is challenging the statistical community for development of less resource-demanding methods for predicting poverty. Traditionally, the proportion of individuals below the poverty line (the 'headcount ratio') is estimated through a fully fledged household budget survey (HBS) covering a period of 12 months and based on diaries or the recall of consumption expenditure on food and non-food items. However, not many countries can justify spending the resources on an annual household budget survey, and consequently proper poverty measures are collected only every fifth or even tenth year. However, annual lower cost 'light surveys' (e.g., CWIQ[1] surveys) are common, and they can be used for predicting poverty. The approach taken is to estimate annual regional/district poverty headcount from the light survey with its corresponding uncertainty, without undertaking a full household budget survey (Wold et al. 2004).

Model Features:

A poverty prediction model, which by combining information from a HBS and a 'light survey', yields:

- Annual headcount rate estimates
- On a regional level
- With estimates of their inaccuracy

The basic idea is to utilize the information in a budget survey to identify a smaller set of household variables (indicators) that can be collected annually between two budget surveys. This is done by estimating a relation that links consumption and poverty to the set of indicators through a statistical model, i.e., by constructing a 'consumption model'. The indicators should be fast to collect and easy to measure. Hence, they may be compiled through so-called light surveys without collecting expenditure data. The information obtained from the light survey and the estimated model is used to predict poverty rates. One such method is developed in "A Statistical Model for Fast and Reliable Measurement of Poverty" (Mathiassen, 2005)[2]. However, as this is a theoretical paper, it may not be sufficient for practical application if one does not have a sound understanding of statistical methodology and the requisite statistical software. The purpose of this paper is therefore to present the steps and procedures for predicting poverty from light surveys in a practical manner. For a formal derivation of the method, see Mathiassen (2005)[3].

Three Main Phases of the Approach:

1. Define a model for the relation between poverty and explanatory variables in the first full household budget survey (HBS1)
2. Include these poverty indicators in a light survey, e.g. the CWIQ
3. When a second full expenditure survey, HBS2, is ready, evaluate the model by including the poverty indicators from the HBS1

The paper is organized as follows. In Section 2, we briefly outline the methodology and the main results, without going into technical/statistical detail. Section 3 is concerned with the preparatory tasks and data requirements. In Section 4, we discuss how one should select the set of potential poverty predictors. In Section 5, we show how to estimate a consumption model. Finally, in Section 6, we show how to predict poverty headcount ratios and estimate the uncertainty of the predictions. The methodology is exemplified using data from Mozambique. The link www.ssb.no/en/int. contains annotated SPSS files for the complete prediction

[1] Core Welfare Indicators Questionnaires, jointly developed by the World Bank with UNDP and UNICEF. These surveys are not designed to measure expenditure or consumption but to obtain indicators of welfare and use of and access to public services.

[2] The challenge to predict poverty is not a new one. Fofack (2000) develops a method for ranking households in a CWIQ survey into expenditure quintiles based on the number of individuals with predicted consumption within each quintile. This method has been applied to, amongst others, Ghana (Fofack, 2000) and Uganda (McKay, 2001).

[3] The methodological approach in this paper is inspired by statistical modeling in the adjacent area of poverty mapping, cf. Elbers, Lanjouw and Lanjouw (2003). The method described in Mathiassen (2005) is based on a simpler approach that enables us to derive closed-form expressions for the standard error of the predictor and also facilitates the statistical estimation. The method presented here, however, also rests on a more stringent assumption that will be discussed and tested.

process, from the preparation of the user files to the estimation of uncertainty in the estimated headcount ratios.

All methods for predicting poverty by applying a consumption model and predicting a future survey critically rely on the assumption that the relation between the consumption variable and the poverty indicators are stable over time. This assumption cannot be tested without two or more budget surveys at hand, or at least a short-form questionnaire on consumption in a light survey. Thus, one should be careful predicting poverty more than a few years into the future or the past, especially in rapidly changing economies, as the relations between the variables are likely to change with the economy.

2. The methodology

In this section, we discuss the methodology for predicting poverty rates with limited reference to the statistical methods. Readers looking for references should consult Mathiassen (2005). Additional formulas needed for the practical application of the method are given in the Appendix of this document.

2.1. A predictor for the headcount ratio

An individual is considered poor if his or her consumption or income falls below a certain threshold. This threshold defines the poverty line. We want to predict the headcount ratio, i.e., the proportion of individuals with consumption below a given poverty line[4].

Let Y_i denote the consumption for individual *i*. We refer to Y_i as household consumption per capita or the adult equivalent. Let *z* denote the poverty line. Let $y_i = 1$ if individual *i* is poor where $Y_i \le z$, and zero otherwise. We are interested in predicting the headcount ratio, y i.e., the share of poor individuals in a population Ω consisting of N^H households. The population can, for example, refer to a region within a country. Because the unit in the survey is the household, one needs to adjust for the number of members in each household. Let s_i be the number of members in household *i*, and let N be the number of individuals in the population. In our case, an individual is considered poor if his or her household's per capita consumption is at, or below the poverty line. Hence:

$$(1) \quad y = \frac{1}{N}\sum_{i\in\Omega} s_i y_i \ .$$

As indicated above, we wish to use a model to predict y for a given set of household variables (indicators). We next assume that:

$$(2) \quad \ln Y_i = X_i\beta + \sigma\varepsilon_i$$

where X_i is the vector of selected poverty indicators, β is a vector of unknown parameters and ε_i is an error term that is assumed to be distributed according to the standard normal distribution. The parameter σ therefore represents the standard deviation of $\sigma\varepsilon_i$. The assumption on normality is, as shown later, used in the step below; however, other distribution functions can be applied. Assume further that ε and X are uncorrelated. In particular, we assume that ε is uncorrelated with household size (or adult equivalents), because household size is used to calculate per capita consumption. The logarithmic transformation of the consumption variable serves to reduce the usual asymmetry in the distribution of the error term and stabilizes the variance. The assumption on homoskedasticity and normality of the error term will be further discussed and tested in the empirical section.

Because of the stochastic component in the estimated consumption level, all individuals have a nonzero probability of being poor[5]. Thus, rather than counting the number of individuals with predicted consumption below the poverty line to find an estimator for the headcount ratio, we use the average probability that an individual is poor as the predictor. The probability that individual i's consumption falls below the poverty line, z, is found by inserting the regression model in a probability function:

$$(3) \quad \begin{aligned} P_i &= P(Y_i < z) = P(\ln Y_i < \ln z) \\ &= P(X_i\beta + \sigma\varepsilon_i < \ln z) = \Phi\left(\frac{\ln z - X_i\beta}{\sigma}\right) \end{aligned}$$

where $\Phi(\ )$ denotes the standard cumulative normal distribution function (but another distribution function could be applied). Note that when an individual's estimated consumption is very low, the probability of being poor is close to one, whereas individuals with very

[4] We will return to the data requirement and definitions of these concepts in the next section.

[5] However, for households with a very high, predicted consumption level, the error term may be so large that the household members' actual consumption theoretically could fall below the poverty line.

high estimated consumption have a probability of being poor close to zero. When the estimated consumption is near the poverty line, the probability of being poor is around one-half.

One predictor for the headcount ratio in (1) is then given by:

$$\hat{P} = \frac{1}{n}\sum_{i\in S} s_i \Phi\left(\frac{\ln z - X_i\hat{\beta}}{\hat{\sigma}}\right). \quad (4)$$

It can be shown that this predictor is biased. Hence, we will use the formula for the unbiased predictor given in (6) in the Appendix. However, for calculating the standard error of the predictor below, it is the simpler predictor in (4) that is used, because using the biased corrected predictor substantially increases the complexity in the calculations, and the error caused by using the unbiased predictor is marginal.

2.2. The standard error of the predictor

The prediction error is the deviation between the poverty level predicted by our model and the actual poverty level in the population. One way to decompose the prediction error is:

$$\begin{aligned}\frac{1}{N}\sum_{i\in\Omega} s_i y_i - \frac{1}{n}\sum_{i\in S} s_i \hat{P}_i = \\ \left[\frac{1}{N}\sum_{i\in\Omega} s_i y_i - \frac{1}{N}\sum_{i\in\Omega} s_i P_i\right] \\ + \left[\frac{1}{N}\sum_{i\in\Omega} s_i P_i - \frac{1}{N}\sum_{i\in\Omega} s_i \hat{P}_i\right] \\ + \left[\frac{1}{N}\sum_{i\in\Omega} s_i \hat{P}_i - \frac{1}{n}\sum_{i\in S} s_i \hat{P}_i\right] \quad .\end{aligned} \quad (5)$$

The first term on the right-hand side in (5) is the difference between the actual and expected population poverty levels. This captures how the headcount ratio in the population deviates from its expected value. This component can be very small when we provide predictions for large samples.

The second term in (5) is the difference between the expected poverty level and the poverty level predicted by the estimated model for the entire population, Ω. This captures uncertainty from the error in the estimate, $\hat{\beta}$.

The last term in (5) is the difference between the predicted poverty level in the population Ω and the predicted poverty level in the sample S. This is the result of uncertainty because S is a finite random sample.

All error components are also affected by the variation of the X-vector in the sample.

The expression of the variance of the error in (5) and the procedure for estimating this variance are described in the Appendix Section 0.

There are other errors that we are not able to measure and that are thus not included in (5). The most critical is stability of the model parameters. Even if the model relation is true at a given time, the regression coefficients may change over time. When the economy changes, the relation between poverty predictors and expenditure may change as well. The more dynamic the economy, and the more time that passes between the surveys, the more likely it is that the model parameters are unstable. To test this assumption, two budget surveys are required to estimate the two consumption models and to test whether the parameters have changed. A short-form measure of consumption could also help to verify the assumption as one could estimate models based on this information and compare the model coefficients.

3. Preparations

3.1. Required features of the household budget survey

The budget survey is used to estimate the consumption model and to calculate the poverty line. Hence, it is the basis for all further work on the poverty estimates. In order to proceed, the following requirements should be met.

- The budget survey should be representative for the entire area for which one is interested in predicting.
- The budget survey should have been conducted 'recently'.
- The budget survey should include nonexpenditure indicators.
- The data quality should be acceptable.

It is particularly important to verify the first of these requirements. For example, the Angolan household survey in 2000 covered only urban areas. One can hardly defend the use of a model estimated based on only urban areas to predict poverty for rural areas. Normally, one would estimate separate models at the rural/urban level, if not the regional level.

Is a recent, high quality household budget survey available?
'Recently' conducted?
Geographical coverage?
Includes non-expenditure indicators?
Acceptable data quality?

Over time, the implicit relations between total expenditure (and hence poverty) and other variables in a household budget survey are subject to change. If this structural relationship has been substantially altered, the estimated parameters may become biased. Moreover, it may be very difficult to assess such biases without having a second budget survey at hand. How fast a budget survey becomes outdated for use in a poverty prediction model depends on the magnitude and speed of changes in the economy. We recommend that 'recent' be interpreted as allowing a maximum time span of five years between the previous budget survey and the 'light' survey.

It is essential that the budget survey also contain nonexpenditure indicators, i.e., items other than standard consumption and expenditure quantities. Because the light survey usually contains no expenditure variables, the nonexpenditure indicators constitute the joint set of indicators that allow the two surveys to be linked. Any household survey contains geographical and other sampling information, as well as vital information about household members. However, it is important that other nonexpenditure variables also be included. These include housing standards, possession of consumer durables, education, and screening questions (yes/no) on consumption and expenditure for various expenditure groups.

Finally, one should also pay attention to the quality of the budget survey data before proceeding with the poverty predictor model. The first step is to read carefully through the survey documentation to obtain an overview of known errors. Second, one should, if possible, contact those responsible for the fieldwork in order to capture any non-documented errors. However, researchers should also make their own assessment of the data quality by checking whether the distributions of the indicators are reasonable. In some budget surveys, fieldwork tools and procedures (like diaries) do not function as well as expected. Hence, many households end up having imputed values on consumption expenditures. Moreover, these problems may frequently be more common among the poor, illiterate and other marginal groups who live in distant locations or in troublesome regions. Thus, one needs to clarify how serious these shortcomings are, and keep in mind when interpreting the results that biases in the initial budget survey may be carried through so as to cause subsequent biases in the headcount predictions from the light survey.

3.2. The expenditure/income concept

Expenditure/consumption rather than income should be used as the welfare indicator upon which the poverty measure is based. First, consumption is likely to be measured more precisely, particularly in poorer economies. In addition, consumption varies less than income, which may fluctuate considerably throughout

the year. A farmer will typically receive the main share of his or her income at harvest time, while the household smoothes consumption over the whole year (see, for example, Johnson et al. 1990). As such, it is generally accepted that consumption provides a more adequate picture of the poor's well-being than other measures.

The next step is to resolve the concept of 'household expenditure'. At its core, it is, of course, all purchases paid for in cash or kind. Likewise, the market value of consumption of one's own produce is included. These flows must subsequently be standardized to cover the same period, usually one year, to capture any seasonal variation. It is also common to include the market value of living in an owned house, the so-called imputed house rent. Along the same line of reasoning, the value of the flow of services rendered by consumer durables may be added[6].

Decide on individual expenditure concept to be used
Per household?
Per capita?
Per adult equivalent (and how to calculate AE?)

When the concept of household expenditure is clear, it is straightforward to compute each household's aggregate annual expenditure. This information will normally be readily available as a single variable. Although the usual measurement unit in the budget survey is the household, poverty is, as with other measures of well-being, essentially an individualistic concept. Derived poverty measures, such as, for example, the 'Headcount Ratio' or the 'Poverty Gap' are thus defined across individuals. The existence of private household goods implies that it would be a mistake to assign the full household expenditure to each individual[7]. One then needs to adjust for the number of members in each household. Individual consumption is then defined as household consumption obtained from the household budget survey, corrected for the number of members in the household. A complicating factor is that there are different ways of calculating individual consumption. The simplest solution is to adjust household aggregate expenditures for the household size by simply dividing it by the number of individuals living in the household[8]. Another approach is to divide aggregate household expenditure by the number of household 'adult equivalents'. When dividing by the number of adult equivalents, one simply applies a system of weights that depend on the size of the household and the age and sex of the individual household members[9]. The problem is that there is no single accepted adult equivalence scale. However, when dealing with subgroups and particular regions, one would normally follow the same procedure as that used for calculating the national poverty rate[10].

The topics discussed above all deal with the content of the concept of individual consumption expenditures. We now raise two issues of a more technical nature. First, the distribution of expenditure (or income) is usually skewed with a long tail to the right (from a few units with very high values). It is thus common practice to transform the variable by taking its logarithm. This gives a more symmetric distribution of the error term, stabilizes the error variance and prevents some observations receiving extreme influence. All of these are beneficial in the estimation. Even after transforming the variables, there may still be outliers[11]. Outliers are candidates for further analytic treatment. One should, as far as possible, check whether outliers are due to errors (in data entry or in use of the questionnaire) and if they are, remove them. In some cases, it may also be necessary to remove other extreme but still correct observations, as they may radically alter the estimated parameters.

As an empirical example, we use IAF 2002/03, the latest household budget survey in Mozambique[12]. Figur 1 and Figure 2 illustrate how the empirical distribution of expenditure per capita in the rural sample changes when one takes the log. The original distribution is skewed (Figure 1). When one applies the log, the distribution appears more symmetrical (Figure 2). The same pattern is prevalent for the urban sample (Figure 3 and Figure 4).

[6] Regardless, some ambiguities usually remain. Should, for example, the consumption of tobacco, alcohol and drugs be included?

[7] A 'private good' is a good where one person's consumption of that good prevents other persons from consuming the good. A typical example of a private good in a household is food. A 'public good' for the household, on the contrary, can be consumed by all household members; e.g., the dwelling's building materials and infrastructure.

[8] This is one important reason for the need of clear definitions about who qualifies as a household member.

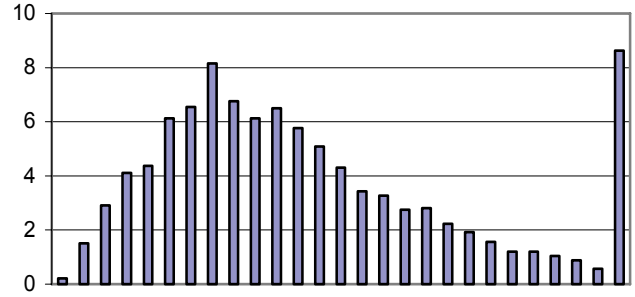
[9] There are two main arguments for using adult equivalents. First, there are economies of scale in household consumption of household public goods. Second, it could be argued that the needs of children are less than those of adults, in particular when food expenditure constitutes a large share of the household's budget.

[10] An underlying assumption for all adjusted expenditure concepts is that every individual household member receives a 'fair share' of the household's consumption of private goods, including, for example, food and clothing. However, qualitative surveys have repeatedly shown that this assumption is violated. Hence, it is reasonable to assume that women and children in nonpoor households are still individually poor and vice versa. In spite of this evidence, we do not consider intra-household distribution effects in this framework because it is very difficult to collect high-quality quantitative data on individual consumption, and because intra-household issues render poverty analysis much more complicated (see Deaton (1997), Chapter 4 for a review).

[11] It is common to define outliers as cases more than three standard deviations from the sample mean.

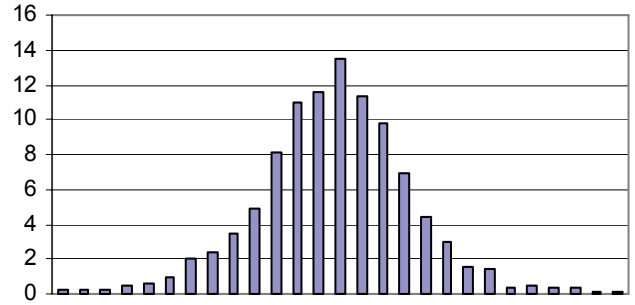
[12] Because this is meant as an illustration, we focus only on one rural region in Central Mozambique.

Figure 1. Distribution of expenditure per capita. Central Rural Mozambique[1]



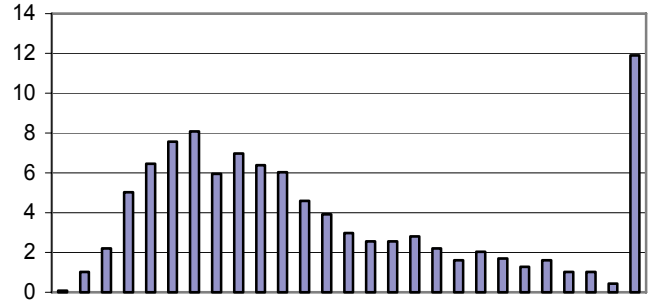
[1] Expenditure per capita is divided into 27 categories, each with an interval of 1,000 Metical (MOM), except for the last category, consisting of all individuals with expenditure above 26,000 MZM.

Figure 2. Distribution of log expenditure per capita. Central Rural Mozambique[1]



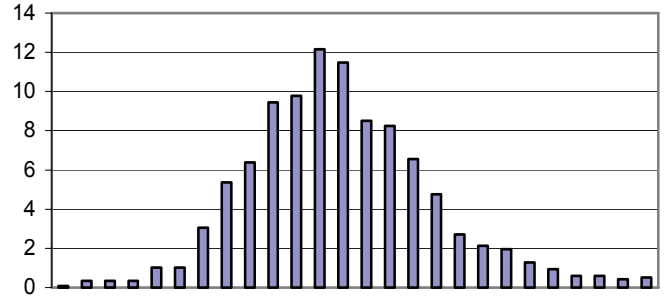
[1] Log expenditure per capita is divided into 27 categories, starting at 6.2 and with an interval of 0.2, up to the last category, consisting of all observations with log expenditure higher than 11.8.

Figure 3. Distribution of expenditure per capita. Central Urban Mozambique[1]



[1] Expenditure per capita is divided into 27 categories, each with an interval of 1,000 Metical (MOM), except for the last category, consisting of all individuals with expenditure above 26,000 MZM.

Figure 4. Distribution of log expenditure per capita. Central Urban Mozambique[1]



[1] Log expenditure per capita is divided into 27 categories, starting at 6.2 and with an interval of 0.2, up to the last category, consisting of all observations with log expenditure higher than 11.8.

3.3. The poverty line

The poverty line is the cut-off point that classifies individuals as poor or non-poor. A national poverty line will normally be available, constructed based on the most recent budget survey. There are two main classes of poverty lines: absolute and relative. In most developing countries, a version of the former is used, often based on the Cost of Basic Needs approach[13]. One first defines the 'food share' of the poverty line as the cost of a minimum calorific intake of a common food basket, considering the average calorie needs of the population. The average cost of this consumption is called the 'food poverty line'.

Decide on poverty line
Monetary based or multi-dimensional?
Absolute or relative?
National, urban, rural, regional?

In the second step, when one adds 'non-food necessities', it usually switches to a relative concept. Common approaches are defining non-food necessities as the average non-food expenditure consumption among households with either total household expenditures, or food household expenditures, around the cost of the food poverty line. Using total household expenditures justifies the use of the concept 'necessities' because these households, in a position where the members can barely be adequately fed, still choose to give up some food consumption in order to consume these non-food goods and services[14]. This also defines the minimum poverty line level, while referring to food household expenditures gives the maximum poverty line level.

Because diets and prices vary, one will often calculate separate poverty lines for urban/rural areas, as well as for regions. For example, staple foods tend to be relatively cheaper, and non-food items relatively more expensive, in rural than in urban areas. The specific content of the average food basket, i.e., the composition of the food items that are used to compute the cost per calorie, often differs between domains. To account for the differences in relative prices, one needs to deflate the prices. This applies in the dimensions of both time and space, as food prices especially tend to vary with both[15]. Prices may be deflated by price level indices or indirectly by calculating separate food poverty lines for the different domains. In Mozambique, the poverty line was constructed using a national food

[13] Mozambique's national poverty line is based on this approach.

[14] See National Directorate of Planning and Budget et al. (2004) for documentation concerning the construction of the poverty line in Mozambique. See Ravallion and Bindani (1994) and Ravallion (1998) for a general discussion on the construction of poverty lines.

[15] Prices in many developing countries rise at a tangible annual rate. Moreover, food prices fluctuate according to the agricultural season, dropping sharply at harvest time. Regional price differences may also be very high in developing countries because of long distances and substandard communication infrastructure.

basket, and there is a single poverty line that has been spatially and temporarily deflated.

3.4. Required features of the 'light survey' (or other survey)

The basic idea of the poverty prediction approach is to use information on the poverty indicators from a 'target survey' to predict per capita expenditure and, in the next instance, the headcount ratio for the target population. We refer to the target survey as a light survey—as it will be in most cases. The concept of a 'light survey' covers a class of household surveys that are less costly and much easier to administer than full-scale budget surveys (Loureiro, Wold and Harris 2006). Hence, light surveys may be conducted more frequently than the large-scale budget surveys, usually on an annual basis. Light surveys usually lack estimates of expenditure, but contain 'sufficient' variables present in the budget survey (we return to what is meant by 'sufficient' later). However, the target survey may also be another budget survey. For example, a change from collecting expenditures using a diary to using a recall approach implies less direct comparability between the aggregate household expenditures in two household budget surveys (see, for example, Tarozzi (2004) for a related method for computing comparable poverty estimates in a similar case). Finally, the target survey may be another budget survey with the comparable poverty estimates used for testing the method.

As discussed, it is important that there be a limited time span between the budget survey and the light survey. Because the poverty predictions critically depend upon the assumptions of stability in the relations between the nonexpenditure indicators associated with poverty and expenditure per capita over time, budget surveys become outdated for such use faster in dynamic, changing economies. It is also important to be aware that the model needs stable relations more than stable variables. In fact, one would expect less stable variables such as screening consumption variables (e.g., yes/no to any meat consumption last week) to ensure a more stable correlation to consumption expenditures than standard household background variables.

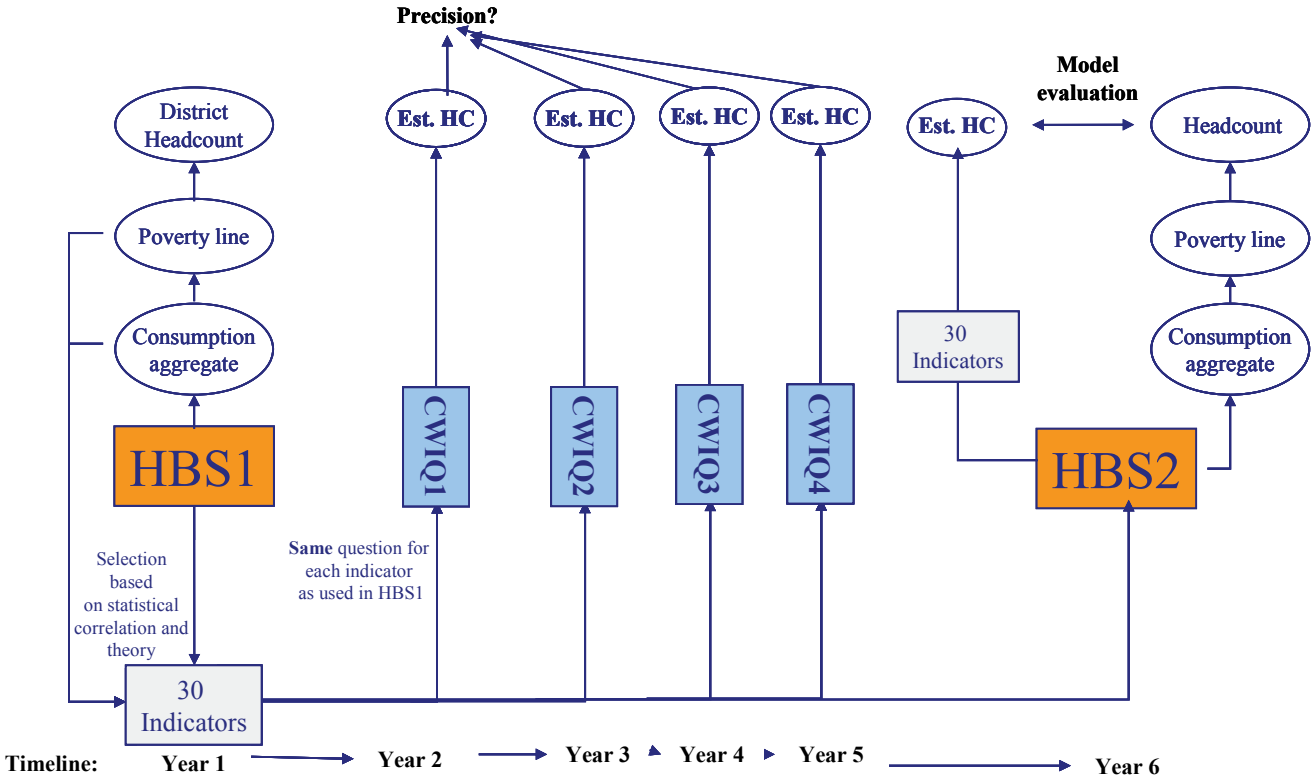
One typically faces one of two different situations.

The first is that a light survey is going to be set, and the indicators to be included for predicting poverty are to be selected. The selection of indicators may then be conducted freely among the variables of the budget survey, for example, by requesting that the light survey administrators add a set of 10–15 questions that would otherwise not have been included. It is critically important that the indicators used for prediction be phrased in exactly the same way as the two surveys.

The second situation is when a light survey, for which one wants to predict the headcount ratio, already exists. In the latter case, all variables in both the expenditure and the light surveys are given. The only available indicators are those that appear to be phrased in exactly the same way in both surveys. We continue to discuss the practical selection of the variable sets for the modelling purpose in the next chapter.

Figure 5 illustrates the entire poverty prediction sequence over a six-year period from the completion of the first budget survey (HBS1) to completion of the second budget survey (HBS2). The light surveys (CWIQ 1–4) are standardized and conducted annually. As discussed above, the model can also be applied to the second budget survey, given that the original indicators are included and the methodological approaches are uniform. This allows for evaluation of the performance of the poverty predictor model. The second budget survey, HBS2, is then used as the base for future predictions of poverty.

Figure 5. The poverty prediction sequence



4. The Consumption Model

Work on the consumption model involves three steps. First, the initial set of indicators is selected based on the criteria described below. Second, the model is estimated. This involves several estimation sequences where different sets of indicators are included and tested. The models are then compared before the model based on the final selection of indicators is chosen. Finally, the assumptions that the model relies on are tested. All of these steps can be performed using only the budget survey dataset, although if a light survey has already been conducted, it must be ensured that the selection of indicators is restricted to those variables common to both surveys.

4.1. Considerations governing the selection of poverty indicators

In this section, we discuss considerations governing the selection of poverty predictors (referring to the selected set of predictors as X). It is crucial for the model's performance to identify good and feasible predictors. The selection of indicators is thus usually the most time-consuming part of the analysis. As discussed in the previous chapter, there are two main situations with respect to the timing of the light survey. In the first situation (I), a light survey is going to be set, while in the second situation (II), a light survey already exists. The approach for the indicator selection differs slightly in these cases.

Two situations for indicator selection

1. A light survey is planned:
   Unconstrained selection from HBS
2. A light survey has already been conducted:
   Constrained selection from HBS

If the light survey is in the pipeline (I), one selects poverty indicators from an unconstrained set of candidate variables that are included in the most recent budget survey. For practical and budgetary reasons, one can, however, only expect to add a limited set of new variables to the light survey (10–15 new variables are usual).

In the case where the light survey has already been conducted (II), one can only select indicators from the set of common variables in the light survey and the budget survey, i.e., indicators derived from questions that are phrased in exactly the same way in both surveys. If this common set is close to empty for a given light survey, that survey cannot be used for predicting poverty. On the positive side, because the light survey has already been conducted, there are no additional costs of adding variables. If need be, one can initially use as many of these common variables as one wants for predicting poverty

In the case of Mozambique, we predicted poverty in early 2006 by combining the budget survey, IAF 2002/03, with the existing labour force survey, IFTRAB 2004/05. As the fieldwork for both surveys has been completed when the analysis was initiated, we had no influence over the choice of variables to be included in the surveys (i.e., situation (II) above).

4.1.1. Criteria for poverty indicators

The first and basic common criterion for situations (I) and (II) is that the poverty indicator candidates be directly available from the Household Budget survey (HBS) questions or can be constructed from them (e.g., the dependency ratio). In the case where the light survey has already been completed, indicators require exactly the same wording, including compatible answer values. In situation (I), one adds new questions to the light survey questionnaire, while making sure that the wording is kept exactly as it was in the household budget survey.

Potential poverty indicators should also be reliable. The reliability criterion implies that one should avoid using as indicators variables that have many missing observations[16]. Moreover, reliability also implies that one should avoid variables that give excessive room for interpretation or subjective assessment among interviewers and/or respondents. These include, for exam-

[16] There is no problem in having variables that are missing due to natural reasons: for example, lacking information about a spouse in households where there is no spouse. We shall return to this later.

ple, subjective assessments of the type: "Do you feel that you are better or worse off now than a year ago?"

The poverty indicators should be:
- Present in the household budget survey
- Based on questions phrased in the same way in the two surveys
- Reliable
- Be "fast" and easy to collect

Finally, where new variables are to be included in a future light survey (case I), these variables must be quick and easy to obtain information about (given the nature of this type of survey). The potential indicators could make a long list, and there are many other considerations to be made along the way, some of which must be done on the basis of subjective considerations. We will use the data from Mozambique to illustrate this process and recommend how to proceed step-by-step[17]. Before this, however, we have some general comments regarding the characteristics of the potential indicators.

4.1.2. Substantive topics and measurement unit of indicators

The Mozambique indicators include variables along the following dimensions describing the welfare of households and their members: Demographic composition, Literacy, Education, Employment, Assets, Dwelling characteristics (type of roof, walls, toilet, number of rooms), Energy and water use, Screening consumption (dichotomous or 'yes/ no' variables only).

If feasible indicators are available, one can also add variables from topics such as Health, Agriculture and Community. In the end, the variables to be included in the analyses depend on the questions requested in the budget survey. Hence, the importance of making sure that the question is phrased in exactly the same way in the two questionnaires cannot be stressed enough[18].

Classification of variables:
- Appear at the individual, household or cluster level?
- Topic/ welfare dimension?
- Stock or flow (volatility)?
- Measurement level?

Another requirement is that all variables eventually appear as indicators at the household level. This is because budget surveys use households as their investigating unit. In the case where individual-level variables form the basis for the indicators, they must be aggregated to the household level. For example, variables measuring the education of individuals can be aggregated into household level indicators, such as the maximum education of any household member, or the education of the most-educated female household member, and so on. The variables sex and age may, in a similar manner, be transformed into the number of adult males, the number of adult females, the number of boys and the number of girls. By combining individual roster information about age, sex and relation to the household head, one may additionally define a 'household type' indicator, taking values such as: 'single person', 'nuclear family without children', 'nuclear family with children' and 'extended family'. Finally, one may also argue that key individual characteristics of the household heads are, in effect, properties of their respective households, and use such individual level information directly as a household level indicator.

Although it is an advantage that the indicators cover different topical dimensions of well-being, the key property of a set of indicators is their ability to predict poverty jointly. A useful approach, regardless of the topical dimension of welfare, is to distinguish between indicators that are expected to be relatively stable over time, and indicators that capture recent changes in the household's situation. Indicators like the maximum education of any household member, the household's ownership of assets and the properties of the dwelling are typical 'stock' variables that change little in the short term, even in households exposed to shocks. For our purposes, they are still useful as a cross-check that the indicators for the budget survey and the light survey are consistent where little change is expected. On the other hand, it is also very useful to include indicators that are able to reflect recent changes and that may help to capture the current situation of the household. Especially in the case of idiosyncratic shocks, such indicators are essential[19]. Typical examples are dichotomous variables of the type: "Did you pay for public transport last month", or the employment status of the household head (or main breadwinner), because these variables may change very quickly, and such changes are likely to be correlated with the household's poverty status.

4.1.3. Continuous, dichotomous or categorical variables

One must also pay close attention to the variables' measurement level, i.e., whether they are continuous,

[17] In the appendix Section 0, we have included the entire list of indicators tested for Mozambique. As indicated, the first part of the list includes indicators included in both the IAF and the IFTRAB. The second part of the list presents the indicators only available in the IAF.

[18] In Mozambique, because the variable "number of rooms in the house" in the IAF was rephrased to "number of rooms used for sleeping" in the IFTRAB, we were unable to use this variable to predict the IFTRAB.

[19] An idiosyncratic shock is a sudden negative event that affects only one or very few households. Typical examples are the death of economically important household members, divorce, prolonged illness, unemployment, etc. On the other hand, shocks like drought, which affects all households in an area, are usually captured by community information.

ordinal, dichotomous or categorical (nominal). Typical continuous ratio variables are variables with many answer categories. Typically, these include 'age', 'asset index score' and the 'literacy ratio'[20]. For ratio variables, expressions such as 'twice as many' are meaningful. Ordinal variables also have ordered answer categories, but one cannot compare the relation between categories as for ratios. A typical example is the 'level of education', (whereas 'years of education' is a continuous ratio variable). Finally, for categorical (or nominal) variables, there is no inherent ranking of categories. A particular case is dichotomous variables that take only two values—usually zero or one. They may be ordinal (like 'sex'), or categorical ("did you consume meat last week?").

The main reason to be concerned with a variable's measurement level is that we wish to include them as independent variables in linear regressions. In order not to violate the preconditions of linear regressions, one can only use continuous or dichotomous independent variables. Non-dichotomous ordinal and categorical (nominal) variables must therefore be transformed into dummy variables. This is accomplished by letting each single answer category of the original variable form the basis for a new, dichotomous dummy variable. Let us take the variable 'Energy used for cooking' as an example. In our dataset, this has the following categories: If cooking with charcoal; If cooking with electricity; If cooking with gas; If cooking with paraffin; If cooking with sawdust; If cooking with wood; If cooking with other energy (unspecified). One could construct new dummy variables for each of the seven categories. The same procedure may be followed for all other variables that are non-dichotomous categorical variables.

Because we are interested in forecasting, rather than the casual relation between each predictor and poverty, we only include significant dummy variables. In this example, we may, for example, use only 'If cooking with wood' in the model. The last category, 'If cooking with other energy' should not be included for further analysis, because it is not clear what the other group contains, and thus it does not fulfil the criteria for being a reliable poverty predictor.

Although we prefer that the answer categories of categorical/nominal and ordinal variables be identical in the expenditure and light survey, they may consist of different numbers of categories, given that one can establish a unique key between them across the two surveys. For example, various types of postsecondary education can (and should) often be collapsed into a single category for 'higher education', although one must always ensure that the content of the variable's other categories are consistent between the surveys.

[20] By definition, the share of those in a certain age group than can read and write varies between zero and one.

4.1.4. Cluster-level variables

Both the expenditure and light surveys are usually based on two-step, 'clustered' sampling designs. First, 300–500 household clusters are selected. Second, 15–20 households are selected from each of the clusters. For some indicators, households tend to be more similar within a cluster than between clusters. Typically, in urban areas, a rich household often lives in quarters with many other rich households, and poor households tend to live together with other poor households[21]. Knowledge about the poverty status of one household thus usually gives a good indication of the poverty status of other households in that cluster, and consequently, the effective sample size of the cluster sample survey decreases compared with a situation of genuine simple random selection. The most highly correlated type of indicators for households (and individuals) in a cluster are community variables, such as distance to the market and the availability of electricity. Including such variables can reduce the effect of clustered variables in the models.

4.1.5. Variables dealing with consumption

Essential features of light surveys are that they do not ask detailed questions about household consumption. Questions about consumption in the budget survey are, in general, comprehensive, as they should include information on the consumption of own production, purchases and gifts. Because it is necessary that the question in the light survey be repeated in exactly the same way as in the budget survey, consumption variables are not fast enough to obtain information about, and in particular, consumption of own produce is not generally considered a 'reliable' variable. However, in the light survey, one may include expenditure variables that are seldom produced in the household, for example, cooking oil and soap. Another potential problem with expenditure variables is that if the information about food consumption in the budget survey questionnaire is based on a diary, rather than on recall, one is not able to reproduce the same interview setting in the light survey, which is based on only one visit[22].

One may, however, include variables that capture simple, dichotomous information on consumption of a semi durable or a list of items. Usually, it is exactly these types of variables that will change rapidly if the household is subject to an idiosyncratic shock. This may substantially increase the explanatory power of the model.

[21] However, this is not always the case. Sometimes, poor squatters live side-by-side with rich households.

[22] In the diary approach, households keep a diary over a certain time in order to record the daily consumption of each item in a list of food items.

4.1.6. How to treat variables that are missing for valid reasons

Some variables contain many missing values as a result of widespread non-response or faulty fieldwork. However, for other variables, some units have missing values for valid reasons. This will be the situation, for example, for variables capturing information about the spouse in a household, when there is no spouse present. Let us label these 'invalid' and 'valid' missing cases, respectively. Variables with many invalid missing values are of little use in regression because they reduce the effective sample size that goes into estimation of the consumption model. If other variables also have many invalid missing cases, but for other units, the aggregate loss of cases may easily become unacceptable.

For valid missing cases, one may, in the 'missing spouse example', simply solve the problem by transforming the missing observations into a 'no spouse in the household' dummy variable. This procedure should be repeated for each original variable that captures a characteristic of the spouse. Let us, for example, assume that a variable concerning the 'years of education' varies between zero and 20. Households with no spouse have a value of zero. In addition, one needs to include a dummy for whether there is a spouse in the household or not. The example can be illustrated with a simple consumption model with only two variables.

$$\ln Y_i = a + bX_i + cEduc_i + \sigma\varepsilon_i$$

Let $X_i = \begin{cases} 1 \; if \;\; no \; spouse \; in \; hh \\ 0 \; if \;\; spouse \; in \; hh \end{cases}$

and $Educ_i = years \; of \;\; education \;\; for \; spouse$.

The model when there is no spouse in the household:

$$\ln Y_i = a + b + \sigma\varepsilon_i ,$$

and correspondingly, if there is a spouse in household, the model is given by:

$$\ln Y_i = a + cEduc_i + \sigma\varepsilon_i .$$

4.1.7. Additional explanatory variables

The list of potential indicators should also include square terms (and possibly log terms) of continuous variables. If one includes log terms, one must make sure that the variables one wishes to transform cannot take the value zero, for which the log function is not defined[23]. Standard budget surveys have their interviews spread out evenly across one year in order to cover seasonal variations adequately. Conversely, light surveys are designed to be as quick as possible, interviewing for only one to three months. When the light survey covers only a part of the year, one could divide the year into, for example, four seasons and include a dummy for each of these in the expenditure model. When predicting, the predicted consumption per capita then has to be adjusted to account for the yearly variation.

Summing up preparations for indicator selection:

Identify the set of feasible variables from the household budget survey:
- Reliable, fast and easy to collect
- Covering various welfare dimensions
- Comprising both stock and flow variables
- Check consistency of stock variables across surveys
- If light survey exists, exactly same questions as in HBS
- Drop if too many cases with illegal missing values

Transform original variables if necessary:
- Aggregate all individual variables to household level
- Generate new, cluster level variables
- Transform all non-continuous variables into dummies
- Transform 'legal' missing cases into new dummies
- Construct additional, grouped or logged variables

Because we are not interested in causality, rather the prediction of poverty, we may also construct more indicators from the same set of original variables than one would do for an analysis of the mechanisms leading to poverty. For example, in addition to using age as a continuous variable, one could also construct age groups for each household and then construct dummy variables for each of these groups. It is most important to attempt to squeeze information from the data when one has fewer candidate variables, as in 'poverty mapping'. Here, a population census corresponds to the light survey. Because the marginal costs of including additional questions in the census questionnaire are very high, one typically has to manage with a very limited set of indicators.

Another special case worth mentioning is when we sometimes include a per person variable for indicators that are essentially private goods: for example, owning a bicycle. In this case, one would construct one variable denoting whether the household owns a bicycle or not, and one variable denoting the number of bicycles per capita in the households.

[23] If that is the case, a trick is to transform the original variable by first adding one to each variable, and thereafter taking the log. This is acceptable because we are not interested in causality, i.e., the value of the estimated parameter, but simply its ability to predict poverty.

4.2. Selection of indicators and estimating the consumption model

Given that the general requirements described in the text box above are taken into account, one wish to identify a final set of poverty indicators containing the variables that jointly have the highest predictive capability of household per capita consumption. This section will illustrate how one identifies this set in practice.

The general approach for the selection of indicators is to compare the estimated models using various combinations of potential poverty indicators as independent variables. In most statistical programs, the process of comparing and evaluating all possible combinations of estimated models is automated and is labelled the stepwise procedure. Based on statistical criteria, the set of indicators that constitute the best model for predicting the poverty headcount ratio in some sense are selected. However, a weakness is that this validates the selection of variables based on the dataset that generated them, and this does not distinguish between actual important predictors and those due to chance alone[24]. Also, because the method (described below) is myopic, looking only one step ahead or backward at a time, one may ignore variable sets that jointly have considerable predictive capability but separately do not. In general, a problem with such automatic methods is that "… they often substitute for thinking about the problem" (Gallard, 2006). Thus, one should be careful when selecting the initial set of variables and should examine the final model carefully to make sure that the selected variables have the expected sign.

In the situation where the variables are added to a pending light survey questionnaire (Case (I) above), there should not be too many indicators. However, one should preferably select enough variables that the marginal gain of including one additional variable is low. The optimal number of new variables to be added to the light survey is, of course, dependent on the particular light survey questionnaire and budget survey data set.

The consumption model should be estimated separately for each region/group. This implies not only that the cases included in the regression equations are mutually exclusive, but also that the models may include different independent variables[25]. The advantage of estimating models at each geographical level is that such a model better captures geographical differences in the economic fabric. For example, it appears essential to estimate separate urban and rural consumption models. However, because the number of observations in a model decreases when we only focus on a subpopulation, the precision of the estimated parameters also decreases. As a rule of thumb, one should ensure that there are at least seven to eight significant independent variables in the selected consumption model.

In our empirical example from the last Mozambique budget survey (IAF02), we estimated seven separate consumption models, each comprising between 800 and 1,900 observations[26]. As a reference, we also estimated a national urban and a national rural model, as well as a full-coverage national model comprising all cases.

Stepwise selection of independent variables:

The forward stepwise regression procedure begins with no variables in the model. For each explanatory variable, the method calculates the F-statistic, reflecting the variable's contribution to the model. Variables are sequentially included according to the magnitude of the F-value. New variables are included as long as they have a p-value lower than some predetermined value. When new variables are included, others may add less information (due to correlation between the indicators) and the program removes a variable if the significance level of its additional contribution falls below a certain level. The backwards selection procedure works in a similar way but starts with all variables included in the model, and then successively remove variables from the model.

In the case of a pending light survey (situation (I) above), the use of separate urban and rural and possibly regional consumption models makes the selection of new questions to be added to the light survey questionnaire a bit more complicated. Eventually, the results from all separate models have to be combined into a common suggestion for variables to be included in the new national light survey questionnaire. In the case of Mozambique, there were about 50 variables that were candidates for inclusion in the light survey questionnaire, based on the criterion that the marginal gain of including each was sufficiently high. However, the selection challenge was simplified by the fact that many of these variables were likely to be included in any future light survey questionnaire anyway, because they were 'standard' household survey variables such as region, gender, household size, and so on. Thus, one should first identify those variables that already are included in the questionnaire. If the remaining list is too long, one should include indicators that appear in many of the models. Also, it is important to include variables that appear in models that have low explanatory power. In general, it is more difficult to find good indicators for rural than for urban areas, and if one has to omit variables, one should omit variables in urban rather than rural models.

[24] See, for instance, Dallal (2006) for a discussion of the shortcomings of the stepwise procedure.

[25] However, the preparatory steps described in the previous section will be the same for all models.

[26] There were three regional-rural models, and four regional-urban models, the latter category including a separate model for the national capital, Maputo.

4.3. Testing modelling assumptions

The parameters for the indicators in the consumption model are estimated with standard Ordinary Least Squares (OLS) regression. Statistical theory asserts that OLS provides the 'best' (unbiased and minimum variance) estimators, given that certain assumptions about the properties of the error term and the independent variables are fulfilled (see Section 2.1). If these assumptions are violated, it will not only reduce the consumption model's performance, but also carry the errors onward into the next step of predicting poverty headcount ratios and their variance. In practical use, however, the assumptions will rarely be completely fulfilled. Hence, it is important to test that possible violations are not severe.

Consumption regression models are often associated with heteroskedasticity. This is a violation of the Ordinary Least Square (OLS) regression assumption that the variance of the error term is constant, regardless of the dependent variable's value. Below, we discuss how one can test this assumption. Second, we test the normality assumption of the error term. This assumption is only required for computing the unbiased estimators for the headcount ratio[27].

4.3.1. Testing for heteroskedasticity

Generally, in the case of expenditure models, the common variance pattern is such that the error term's variance increases with per capita consumption expenditure. One explanation is that it is easier to remember how much one spent if one spent little. Even when heteroskedasticity is present, the parameter estimates for the predictors in the consumption model will still be unbiased, as long as the remaining assumptions are fulfilled. However, special caution should be applied when calculating their standard errors.

Tests of the assumption of constant variance:

The Breusch Pagan and the White test tests the assumption that the variance is not being a function of the explanatory variables.

Visual diagnostics that are used for testing for heteroskedasticity:

Plots of the predicted consumption against:
- The residual
- The squared residual
- The absolute value of the residual

Following, we present an empirical example on how to test whether the model's assumption about constant variance is violated, i.e., that heteroskedasticity does exist. The log transformation of the expenditure variable ensures, as shown above, that we obtain a more symmetric distribution of the dependent variable, and thus also has the potential of reducing the prevalence of heteroskedasticity in the error term. In Figure 1 to Figure 4, we observed how the log transformation of the dependent variable, in the case of the Mozambique IAF budget survey, transformed the distribution into an approximately normal distribution.

Formal tests of the assumption about constant variance may be ambiguous. In the national urban Mozambique model, none of the tests described was able to reject the null hypothesis that the variance of the error term is constant (i.e., that we have the desired homoskedasticity) at the 5 percent level of significance (see Table 4[28]). In the national rural Mozambique model, the Breusch–Pagan and White tests reject the hypothesis about homoskedasticity (see Table 3). These tests are, however, sensitive to the number of observations. As is usual with a large number of observations, even small deviations lead to rejection of the hypothesis[29].

Alternatives to these formal tests include study of the plots of the residual. Below, we show plots of the residual against the estimated conditional expectation (the predicted value). One should expect that if the error term is homoskedastic, the plot would show a random pattern across the entire range of the predicted value. Figur 6 and Figure 7 show the residual versus predicted value for respectively rural and urban. The plots indicate that the assumptions about constant variance (i.e., homoskedasticity) are reasonable.

[27] The test will also determine whether we apply the Normal distribution function in calculating the probability of being poor. If normality is violated, one should examine plots of the residual to determine which distribution function to apply when calculating the probabilities.

[28] The Breusch–Pagan and White's tests consider the general (unrestricted) alternative hypothesis in which no assumptions are made on the residual variances. The Breusch–Pagan tests whether the variance is a linear function of the explanatory variables or function of linear combination of the variables, by regressing the squared residuals on a linear combination of the explanatory variables. The White test is a special case of the Breusch–Pagan test, where the squared residuals are regressed on the explanatory variables, their squares and cross-products.

[29] When we estimate the model using a smaller randomly drawn sample of the expenditure survey (for example 500 rather than 1,900 observations), none of the homoskedasticity tests are rejected.

Figure 6. Plot of residual versus predicted value for Central Rural

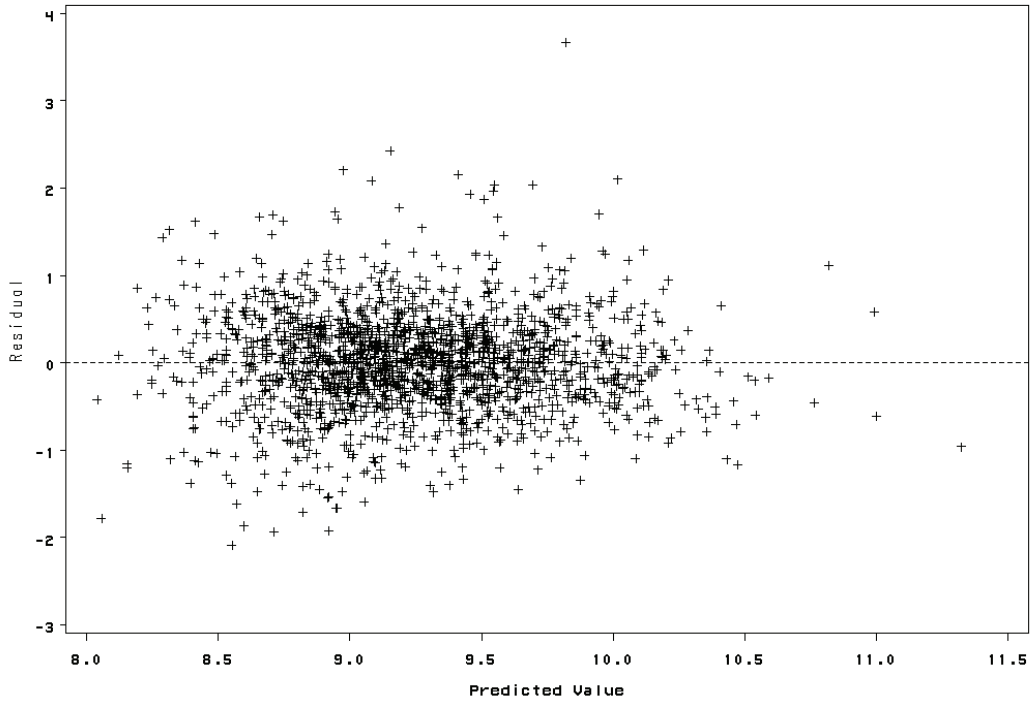
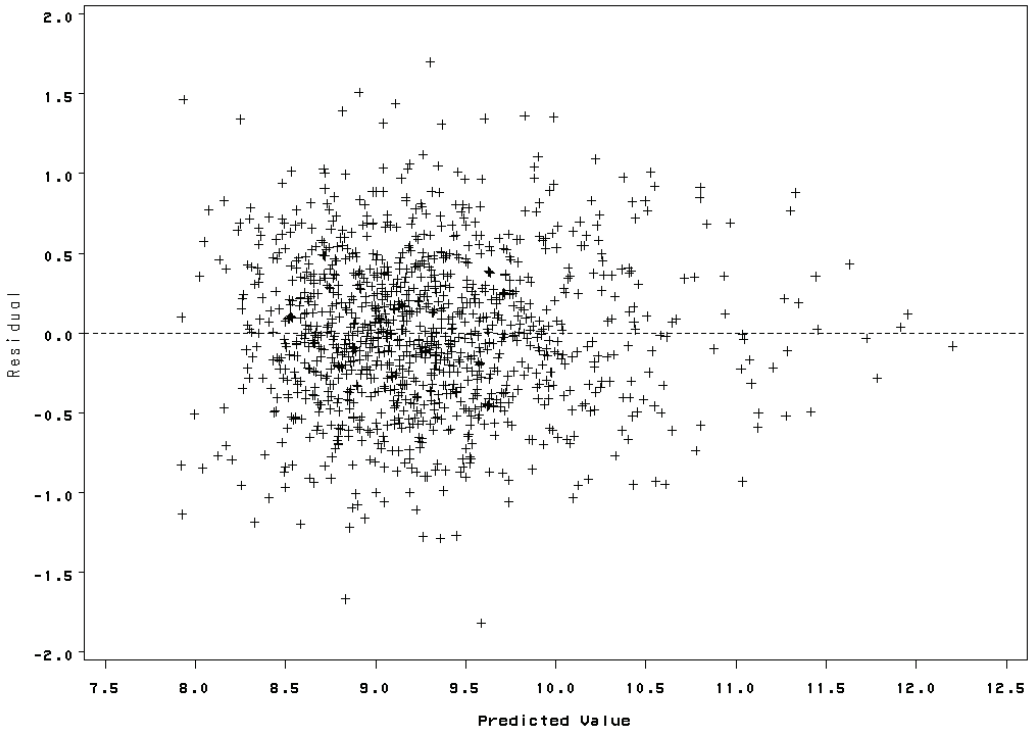
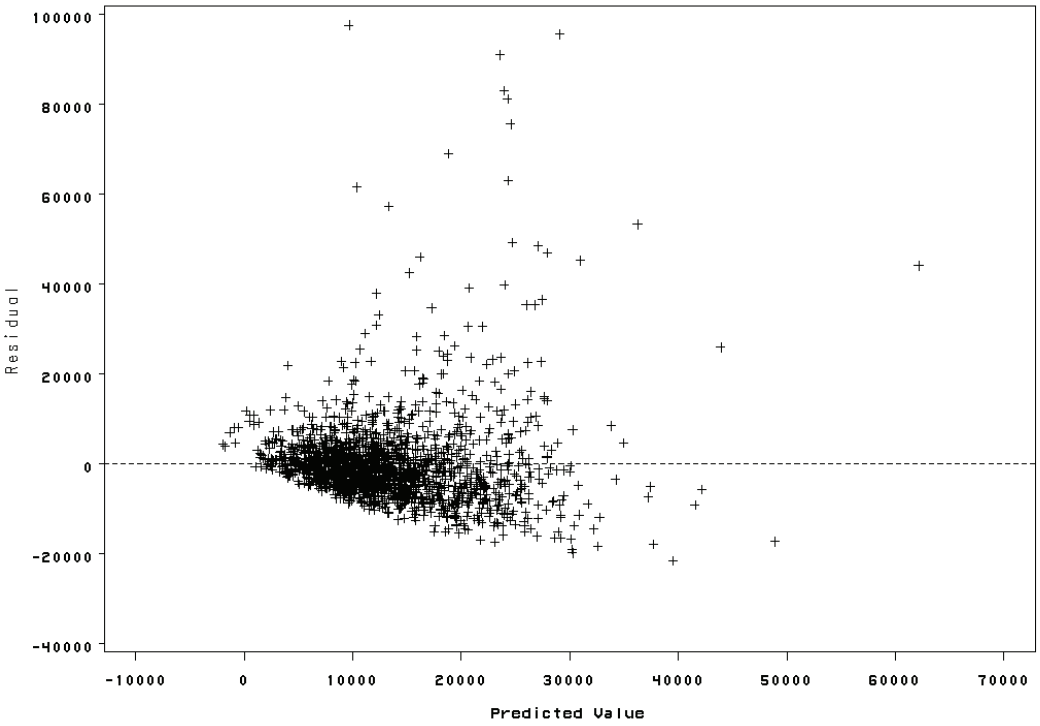


Figure 7. Plot of residual versus predicted value for Central Urban



By comparison, the plot in Figure 1 shows the residual versus the predicted value for the rural Mozambique model when the dependent variable is the untransformed expenditure per capita rather than the log transformation of the expenditure variable. It is fairly evident that this plot indicates the presence of heteroskedacticity.

Figure 8. Plot of residual versus predicted value for Central Rural, when expenditure per capita is the dependent variable



4.3.2. Testing for non-normally distributed error terms

In the computation of the poverty ratio, we need to make an assumption about the distribution function to calculate the individual's probability of being poor. Thus, we test whether it is reasonable to assume that the error terms are normally distributed. Formal tests[30] do not reject the hypotheses about non-normality in the distribution of the error term when all observations used for estimating the consumption model are included, but they are rejected with a randomly drawn smaller number of observations. The same argument applies when using formal tests to check for normality: with large samples, small deviations lead to rejection of the hypothesis of normality. Thus, we also evaluate this assumption by assessing plots. Cumulative normal probability plots (PP-plot) are commonly used for testing the normally distributed error term assumption. A PP-plot compares the cumulative rank ordered values of a variable with the cumulative expected normal values, given the sample size, mean and standard deviation of the variable. The normality assumption is violated if there are serious departures from a straight-line pattern. The PP-plots from both the rural and the urban models seem to be acceptable in this way (see Figur 9 and figure 10).

However, if the normal distribution function is rejected, one should study the empirical distribution of the residual from this one test and apply a suitable distribution function.

Figure 9. PP-plot for residual, Central Rural

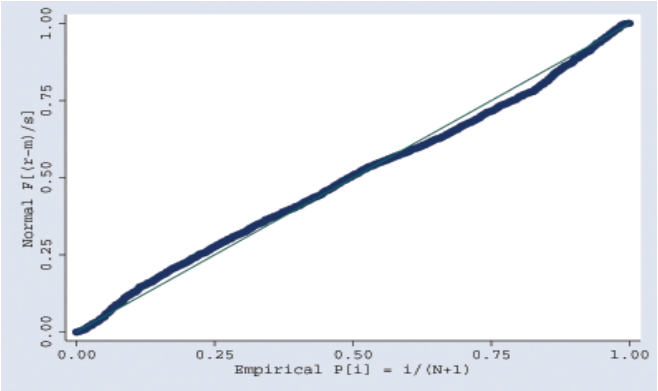
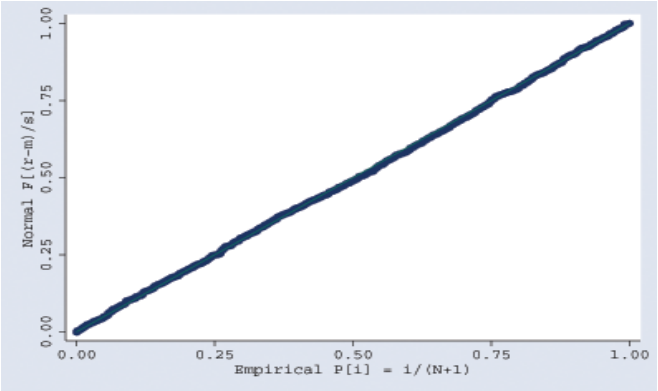


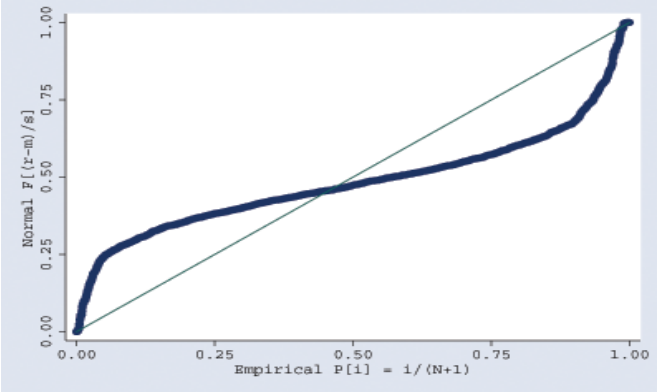
Figure 10. PP-plot for residual, Central Urban



[30] The Kolmogorov–Smirnov, the Cramer–von Mises and the Anderson–Darling tests.

By comparison, Figur 11 shows the PP-plot when expenditure per capita rather than log of expenditure per capita is used as the dependent variable[31]. In this case, one obviously has to reject the hypothesis about a normally distributed error term. Hence, the log transformation of the expenditure variable has been beneficial toward having the regression equation fulfilling the standard OLS assumptions.

Figure 11. PP-plot for residual, Central Rural, when expenditure per capita is dependent variable



[31] The explanatory variables are selected in the same way as when applying the log-linear model.

5. Predicting poverty based on information from a light survey

If the variables that form the basis of the most desired poverty predictors are not already readily available in the light survey questionnaire, one needs, if possible, to include them. It is, as emphasized earlier, of utmost important that the questions be phrased exactly in the same way as in the budget survey and that the reference period be the same. Preferably, the timing of the fieldwork should also correspond, if some of the important predictors are subject to seasonal variations. If this is not possible, it is essential to adjust for the seasonal variation.

The predictor for the headcount ratio is the average probability over all individuals of being poor, as given in equation (8). This probability must be weighted by the sampling weights. Thus, one needs to insert the estimated parameters as well as the poverty line and the predictors to calculate the probability that an individual is poor. For comparison purposes, one should also calculate the headcount ratio based on the prediction within the budget survey sample, in this case the IAF. This serves as a better comparison with the predicted headcount ratio as it is based on the same model, while the actual headcount ratio is calculated based on expenditure. Within-sample prediction is also, to some extent, a test of the model. A large deviation from the actual headcount ratio indicates that the model is weak.

The standard error is simulated according to equation (11) in the Appendix Section 0. One generates 1,000 random draws to compute sufficient variation in the predicted probabilities of being poor and then computes the three different components of the standard error separately. One component is the variation that stems from calculating the expected poverty level rather than the actual poverty level. This component will be very small when predicting for large populations as here[32]. Second, we compute the variance that is due to uncertainty in the model parameters. Finally, we compute the variance due to the sampling uncertainty in the light survey. The total variance of the predictor is found by summing these components.

In the next section, we discuss the results from the predictions as well as the regression results.

[32] As opposed to poverty mapping predicting at a low level, this variance component constitutes a significant part of the total variance.

6. Discussion of results

Table 1 presents the predicted headcount ratios with their standard errors in parentheses; all standard errors are corrected for the sampling design. The fourth column shows the actual poverty level, calculated from the expenditure information in the IAF2002/03. The poverty headcounts for the Central region were 45 and 46 percent in rural and urban areas, respectively. Column five contains the within-sample predictions, which are near the actual poverty estimates[33]. The final column shows the same predictions based on IFTRAB 2004/05 and indicates that poverty fell by 6 percentage points in both rural and urban areas during the two-year period. This observed difference in poverty is, however, not statistically significant.

Table 1. Headcount, predicted and actual with standard errors

	Number of observations		Actual poverty level, expenditure survey	Predicted poverty level, expenditure survey	Predicted poverty level, labour force survey
	Expenditure survey	Labour force survey			
Rural	1924	3535	45.2 (2.9)	47.4	40.9 (3.0)
Urban	1176	2853	46.7 (3.1)	45.5	40.1 (3.1)

*1,000 random draws were used in the simulations

The sampling errors of the actual poverty rates predicted based on the IAF are about the same as the standard errors of the model prediction. Because the light survey IFTRAB consists of a larger sample than the IAF, we have also estimated the standard deviation based on a sample at the same size as the budget survey[34]. Comparing these numbers with the standard errors of the actual prediction in column 3 gives a picture of the effect on the standard error of using a model approach compared with a fully fledged budget survey. The standard error of the actual poverty headcount for Rural Central is 2.9, compared with 3.3 for the standard error of the model-based prediction of the same sample size. The corresponding figures are 3.1 and 3.4 for the urban sample.

The model standard error consists of three components: idiosyncratic errors, sampling errors from the light survey and errors in estimation of the model's parameters. For a sample of this size, the last component is the main contributor to the overall prediction error (contributing about 80 percent when using the full IFTRAB samples). The error because the expected poverty level for the entire population (idiosyncratic errors) differs from the actual poverty level is very small when the population for which we predict is large.

Thus, as the standard errors of the actual predictor consist only of a sampling component, the standard error in the model prediction is much lower than the sampling error of the actual prediction; compare 1.4 and 1.6 for rural and urban, respectively, with 2.9 and 3.1 for the actual poverty estimates. This is because by utilizing a model, much prior information about the expected expenditure level is already given: fewer observations are required to obtain the same precision level.

Table 3 and Table 4 in the Appendix Section 0 present the regression results for rural and urban areas. Among the household-level variables, the common variables in the two models are the asset index and the household size. In addition, dummies for individual assets are included in both models[35].

[33] The within-sample prediction is included as it is directly comparable to the out-of-sample prediction, and it also provides some indication of how the model performs.

[34] We did this by estimating the variance of the predictor from a randomly drawn sample of the light survey, repeated this procedure 50 times and computed the average standard error.

[35] From the same tables, we note that some of the cluster variables have an unexpected sign. For example, if one lives in a cluster where it is common to have a telephone, one tends to be worse off, everything else being equal. We interpret this to reflect that the cluster variables for the share having a telephone can capture effects of other economic dimensions, e.g., areas where a high proportion of the households have telephones tend to be more wealthy areas and thus may therefore also be more expensive. Other cluster variables, however, like education and improved water, have the expected positive effect on welfare in the cluster. It is most likely that these cluster variables jointly capture both dimensions: that better-off clusters tend to have more and more expensive consumption patterns (and thus, each calorie is more expensive), as well as that households located in better-off clusters tend to be better off, everything else being equal.

We have also estimated the consumption model allowing for the final set of indicators to be selected by the model, i.e., the so-called unconstrained model above. Table 2 includes the rural and urban R-squares comparing the 'constrained' with the 'unconstrained' models. We see that R-squared increases by 7 and 6 percentage points in the rural and urban models, respectively. Thus, expanding the possible set of variables to include all desired poverty indicators (see Appendix Section 0) is particularly important for the rural model, because it generally proves to have relatively low explanatory power compared with the urban model.

Table 2. Adjusted R-squared for the models

	IFTRAB	QUIBB
Rural	0,39	0,48
Urban	0,62	0,69

7. Concluding remarks

In this paper, we have outlined step-by-step a method for predicting the headcount ratio and its standard errors. The method can be applied in years when no budget survey other than a light survey is available, or it can be used to produce comparable poverty estimates when the aggregated consumption estimates from separate budget surveys are not comparable because of changes in the framing of the surveys. As discussed, many items need to be considered to ensure that the model predictions are as reliable as possible. However, the fundamental assumption about stable model parameters is often not testable at the time of the light survey. Hence, this assumption should be tested when a new budget survey is available, and generally one should not forecast a long time ahead or backward.

The empirical example in this paper shows that the assumption about homoskedasticity does not appear to pose a problem for the analysis[36], and if one accepts the assumption about stability of the model parameters, the data from Mozambique illustrate that the additional uncertainty from the proposed method is acceptable. The standard errors of the predictions based on the model are higher than the standard errors of the poverty headcounts estimated based on the fully fledged budget survey. It is, however, only about 15 and 30 percent higher in the rural and urban samples, respectively, given the same sample size in the light survey and the budget survey. As the light surveys often tend to be larger than the budget survey, as in this example, the difference in the precision of the predictors is even smaller. Thus, the inaccuracy in poverty predictions based on either a survey or a model is a strong argument for repeated surveys. It also illustrates the possibility of establishing trend statistics rather than using just two surveys with, say, five years between them.

The underlying assumptions, however, have to be tested and evaluated for every analysis undertaken. Refer to Mathiassen (2005) for a procedure if one finds that heteroskedasticity is a problem.

[36] Neither does it seem to be a serious deviation from these assumptions in the analyses of the remaining regions in Mozambique and in corresponding analysis from Malawi based on the Integrated Household Survey 2004, IHS2004.

Appendix

1. Methodological appendix

In this section, we present results of the mathematical derivations of the bias and the standard error of the predictor. The reader may wish to consult Mathiassen (2005) for further details, as well as Green (2003) and Wooldridge (2002) for an understanding of the econometrics used.

we show the It can be shown that an unbiased predictor for predicting the headcount ratio is given by:

(6) $$\hat{P} = \frac{1}{n}\sum_{i\in S} s_i \Phi\left(\frac{\ln z - X_i\hat{\beta}}{\hat{\sigma}\sqrt{\tau_i^2+1}}\right)$$

where:

(7) $$\tau_i^2 = \mathrm{var}\left(\frac{X_i\hat{\beta}}{\sigma}\,\middle|\, X_i\right) = X_i\left(\tilde{X}'\tilde{X}\right)^{-1}X_i'$$

and $\tilde{X}$ is the matrix of poverty indicators obtained from the budget survey given by $\tilde{X}' = \left(\tilde{X}_1', \tilde{X}_2', ..., \tilde{X}_n'\right)$, and X is the matrix given by $X' = \left(X_1', X_2', ..., X_n'\right)$.

Let w_i denote the sampling weight for household i. The predictor is then given by:

(8) $$\hat{P} = \frac{1}{\sum_i w_i}\sum_{i\in S} w_i s_i \Phi\left(\frac{\ln z - X_i\hat{\beta}}{\hat{\sigma}\sqrt{\tau_i^2+1}}\right).$$

It can be shown that the variance of the error in (5) can be written as follows (see Mathiassen (2005)).

(9)
$$\mathrm{var}\left(\frac{1}{N}\sum_{i\in\Omega} s_i y_i - \frac{1}{n}\sum_{i\in S} s_i \hat{P}_i\right) =$$
$$\left(\frac{1}{N}\right)^2 \sum_{i\in\Omega} s_i^2\left(P_i - P_i^2\right) + \mathrm{var}\left(\frac{1}{N}\sum_{i\in\Omega} s_i\left(P_i - \hat{P}_i\right)\right) + \left(1 - \frac{n^H}{N^H}\right)\frac{n^H}{n^2} E\,\mathrm{var}\left(s_i\hat{P}_i\,\middle|\,\hat{\beta}\right)$$

Here N^H denotes the number of households in the target population.

In this expression, we have assumed simple random sampling. We can, however, allow for other sampling designs by adjusting the last term of the right-hand side of
(9), and we will shortly return to how this should be done.

One can use Monte Carlo simulations to estimate the variance given in
(9). It can be shown that one can generate random draws and compute a predictor as follows. Let:

(10) $$D_{ij} = \Phi\left(\frac{\ln z - X_i\beta}{\sigma} - \tau_i\eta_{ij}\right), \qquad \overline{D}_{i.} = \frac{1}{M}\sum_{j=1}^{M} D_{ij}\,, \qquad \overline{D}_{j.} = \frac{1}{n^H}\sum_{i\in S} s_i D_{ij}$$

where η_{ij}, j = 1,2,…M, is i.i.d. random draws from N(0,1). τ_i is given in

(7). Here, D_{ij} is analogue to $\hat{P}_i$ in (5) and corresponds to the j[th] random draw of the stochastic error term. In other words, for each household with the given characteristics, X_i, we generate M independent probabilities of being poor. We use the average over these M simulated probabilities of being poor, $\overline{D}_{i.}$, as an estimator for $P_{i.}$ when computing the variance. By generating random draws, we are able to produce an estimate for the variance of the predictor, even though we initially only had one observation for each individual.

By means of $\{D_{ij}\}$, one can simulate:

$$\frac{1}{N}\sum_{i\in\Omega} s_i^2\left(P_i - P_i^2\right) \quad \text{by} \quad \frac{1}{n}\sum_{i\in S} s_i^2\left(\overline{D}_{i.} - \overline{D}_{i.}^2\right)$$

$$\operatorname{var}\left(\frac{1}{N}\sum_{i\in\Omega} s_i\left(P_i - \hat{P}_i\right)\right) \quad \text{by} \quad \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{n}\sum_{i\in S} s_i\left(\overline{D}_{i.} - D_{ij}\right)\right)^2$$

and:

$$E\operatorname{var}\left(s_i\hat{P}_i\middle|\hat{\beta}\right) \quad \text{by} \quad \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{n^H}\sum_{i\in S}\left(s_i D_{ij} - \frac{1}{n^H}\sum_{i\in S} s_i D_{ij}\right)^2\right).$$

Thus, total variance of the prediction error can be simulated by:

$$\begin{aligned}
&\frac{1}{N}\frac{1}{n}\sum_{i\in S} s_i^2\left(\overline{D}_{i.} - \overline{D}_{i.}^2\right) + \\
&\frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{n}\sum_{i\in S} s_i\left(\overline{D}_{i.} - D_{ij}\right)\right)^2 + \\
&\left(1 - \frac{n^H}{N^H}\right)\frac{n^H}{n^2}\frac{1}{M}\frac{1}{n^H}\sum_{j=1}^{M}\sum_{i\in S}\left(s_i D_{ij} - \frac{1}{n^H}\sum_{i\in S} s_i D_{ij}\right)^2 \quad .
\end{aligned} \tag{11}$$

In the first term, equation
(11), because of the idiosyncratic component, we replace the expected poverty level for each individual with the mean predicted probability of being poor generated by the random draws. We use the variation within the sample n^H as a proxy for the variation within the population.

The second term, because of uncertainty in the estimated model parameters, is the variance of the mean error in prediction. Because we only have predictions for the sample and not the entire population, we use the mean error in the sub sample n^H as a proxy to calculate this variance. We calculate the mean prediction in the sample for each random draw and use these to calculate an empirical variance.

The third term, because of sampling, is the expected variance of the predictor given the estimated parameters. It is computed by calculating the empirical variance of the predictor in the sample and over the random draw. The latter takes care of the fact that it is an estimate for the expected variance.

In the case where we do not have a simple random sample frame, the third term of
(11) can be estimated by using the syntax for estimating sampling variances as given in the packages, for example, SPSS, SAS or STATA. In this case, one specifies Dij as the variable for which one wants to calculate the sampling errors and the strata, clusters and household weights as given by the survey.

2. List of poverty indicators

Common indicators in IAF and IFTRAB

Literacy:
- All adults illiterate
- Some adults illiterate
- One adult illiterate
- No adult illiterate
- Number of illiterate adults in household
- Head illiterate

Education:
- Education of most-educated female member
- Education of most-educated household member
- Education of most-educated male member

Employment:
- Head employed in primary sector
- Head employed in secondary sector
- Head employed in tertiary sector
- If head not employed

Assets:
- Simple additive asset index
- Simple additive expensive asset index
- Beds per person
- Bicycles per person
- Mobiles per person
- Radios per person
- Household owns air conditioner
- Household owns bed
- Household owns bicycle
- Household owns car
- Household owns oven
- Household owns computer
- Household owns electric iron
- Household owns fan
- Household owns freezer
- Household owns fridge
- Household owns hi-fi set
- Household owns mobile phone
- Household owns motorcycle
- Household owns printer
- Household owns radio
- Household owns sewing machine
- Household owns telephone
- Household owns TV
- Household owns wall watch
- Household owns washing machine

Energy, water and sanitation:
- Type of energy used for cooking
- Type of energy used for lighting
- Type of water source
- Type of toilet

Housing:
- Type of roof
- Type of toilet
- Type of walls

Demographic composition:

- Demographic dependency ratio
- Number of members in household
- Number of members younger than 15 years
- Number of persons 65 years or older
- Number of adults in household
- Number of handicapped in household
- Number of daughters of head or spouse in household
- Number of sons of head or spouse in household
- Number of children of head or spouse in household
- Number of spouses in household
- Number of non-relatives in household
- Number of non-close relatives in household
- Number of heads and spouses in household
- Age of household head
- If head is divorced/separated
- If male head
- If head is married
- If head never married
- If head is widowed
- One or two generations with children younger than 15
- One or two generations with no children younger than 15
- Three generations or complex
- Single person
- Single parent with children younger than 15
- Single parent with adult sons/daughters
- Couple with children younger than 15
- Couple with adult sons/daughters
- Couple
- Extended family (outside core)

Indicators available in IAF but not in IFTRAB

- Acquired agricultural tools or inputs last 3 months
- Acquired building materials last month
- Acquired building materials last 3 months
- Acquired clothes or shoes last month
- Acquired clothes or shoes last 3 months
- Acquired domestic utensils last 3 months
- Acquired furniture last month
- Acquired furniture last 3 months
- Acquired soap last month
- Consumed bread last week
- Consumed eggs last week
- Consumed maize flour last week
- Consumed meat last week
- Consumed milk products last week
- Consumed cooking oil last week
- Consumed rice last week
- Consumed seafood last week
- Consumed sweet potato last week
- If no meals yesterday
- If one meal yesterday
- If two meals yesterday
- If three meals yesterday
- If paid for transport last month
- If usually use detergent for washing clothes
- If any household members contracted labourers last season
- If any household members did occasional agricultural work last season
- If household owns poultry
- Rooms per capita

3. Estimation results

Table 3. Regression results, Central Rural Mozambique (OLS)

# Observations:	1917
Root Mean Square Error:	0.56239
Adj. R-Square:	0.3861

Variable	Parameter Estimate	Standard Error	t-value
Intercept	9.8	0.07	140.2
Asset index	0.07	0.01	7.8
No. of spouses in household	0.07	0.02	2.8
If head never married	0.38	0.15	2.6
No. of members in household	–0.21	0.01	–15.0
No. of members in household, squared	0.01	0.00	8.8
If head works in tertiary sector	0.11	0.05	2.4
One or two generations and no children	0.21	0.05	4.4
If household owns bicycle	0.08	0.03	2.6
If household owns hi-fi set	0.29	0.06	5.0
If improved water	0.05	0.02	2.6
If paraffin used for lighting	0.10	0.03	3.6
If no toilet	–0.10	0.03	–2.9
Tete (Province)	–0.22	0.04	–5.9
Sofala (Province)	0.38	0.04	9.6

The regression also includes dummies for agro-ecological zones.
Tests of heteroskedasticity, Ho: Constant variance

White's general test statistic:	340.2	P-value = $5.7e^{-10}$
Breusch–Pagan / Cook–Weisberg test statistic:	8.22	Prob > chi2 = 0.0041

Table 4. Regression results, Central Urban Mozambique (OLS)

# Observations:	1172			
Root Mean Square Error:	0.48997			
Adj. R-Square:	0.619			
Variable		**Parameter Estimate**	**Standard Error**	**t-value**
Intercept		9.90	0.07	144.9
Members in household		–0.24	0.02	–14.3
Members in household, squared		0.01	0.001	7.7
Asset index		0.10	0.01	12.5
Asset index, squared		0.00	0.00	–5.3
If cooking with charcoal		0.15	0.03	4.5
If cooking with electricity		0.47	0.10	4.6
If cooking with paraffin		0.67	0.22	3.0
One or two generations, with children below 15		–0.08	0.03	–2.5
Single person		0.22	0.09	2.6
No. of non-relatives in household		0.15	0.06	2.4
If household owns car		0.55	0.09	6.3
If household owns TV		0.17	0.06	3.0
If candle used for lighting		0.20	0.08	2.4
If asbestos roof		0.08	0.03	2.5
if grass roof		–0.08	0.03	–2.8
One adult illiterate		–0.09	0.03	–2.9
Mobile phones per person		0.48	0.19	2.6
Tete (Province)		–0.33	0.04	–8.3

Tests of heteroskedasticity. Ho: Constant variance

White's general test statistic:	241.8802	P-value = 0.0536
Breusch–Pagan / Cook–Weisberg test statistics	0.27	Prob > chi2 = 0.6066

References

Dallal, Gerard E. "Little Handbook of Statistical Practice" 10/09/2006 <http://www.tufts.edu/~gdallal/simplify.htm> Accessed 10/09/2006.

Deaton, A. (1997): *The Analysis of Household Surveys. A Micro econometric Approach to Development Policy,* The Johns Hopkins University Press, USA.

Elbers, C., Lanjouw, J.O. and Lanjouw, P. (2003): Micro Level Estimation of Poverty and Inequality, *Econometrica*, Vol. 71, No. 1. pp. 355–364.

Fofack, H. (2000): Combining Light Monitoring Surveys with Integrated Surveys to Improve Targeting for Poverty Reduction: The Case of Ghana, *The World Bank Economic Review*, Vol. 14 No. 1 pp. 195–219.

Greene, W. (2003): *Econometric Analysis* Prentice Hall, Englewood Cliffs.

Johnson, M., McKay, A. and Round, J. (1990): Income and Expenditure in a System of Household Accounts: Concepts and Estimation, *Social Dimensions of Adjustment Working Paper* No. 10, The World Bank.

Loureiro, Wold and Harris (2006): Compendium from Workshop on Light Core Surveys for Policy Monitoring of National PRSPs and MDGs in Maputo, December 2005, Documents 2006/9, Statistics Norway.

Mathiassen, A. (2005): A simple poverty predictor model with assessment of the uncertainty. Revised version, Discussion Paper No. 415, Statistics Norway.

McKay, A. (2001): Report on Investigating the Development of a Poverty Correlates Model for Uganda, Unpublished paper.

National Directorate of Planning and Budget, Economic Research Bureau, International Food Policy Research Institute and Purdue University (2004): Poverty and Well-being in Mozambique: The Second National Assessment.

Ravallion, M. (1998): Poverty Lines in Theory and Practice, LSMS Working Paper No. 133, The World Bank.

Ravallion, M. and Bindani, B. (1994): How Robust is a Poverty Line?, *World Bank Economic Review*, Vol. 8, pp. 75–102.

Tarozzi, A. (2004): Calculating Comparable Statistics from Incomparable Surveys, with an Application to Poverty in India, Working Paper, Duke University.

Wold et al. (2004): A Sustainable Household Survey Based Poverty Monitoring System. A Poverty Monitoring System Based upon Household Survey Estimation of Total Consumption. A Preliminary Paper Asking for Cooperation, Documents 2004/17, Statistics Norway.

Wooldridge, J.M. (2002): *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Massachusetts.